Correlated Motions Analysis from Molecular Dynamics Trajectories: Statistical Accuracy on the Determination of Canonical Correlation Coefficients

D. GENEST

CBM, UPR 4301 CNRS, Affiliated to the University of Orleans, rue Charles Sadron, 45071 Orleans cedex 02, France

Received 22 January 1999; accepted 15 June 1999

ABSTRACT: A numerical study of the accuracy on the determination of a unique global canonical correlation coefficient C between two groups of random variables is presented as a function of the number of variables of both groups (n and m, respectively), of the sampling size N and of the actual level of correlation C between the groups. The method used to estimate C has been already described (Briki, F.; Genest, D. Biophys Chem 1994, 52, 35–43; and J Biomol Struct Dynam 1995, 12, 1063–1082), and is implemented in the home made program TECOR. To check the accuracy on the estimation of C for given values of n, m, N, and C, samples of the random variables are synthesized (with known C), then TECOR is used to get an estimated value M of the global canonical coefficient, which is compared to the actual value C. Special attention is paid to the application of the method to the analysis from molecular dynamics simulation trajectories of concerted motions of two groups of atoms (not larger than about 20 atoms) in the course of internal deformation of biopolymers. It is found that there is a good agreement between M and C for moderate and high correlation ($C \geq 0.35$), provided that at least about 2000 configurations are stored during the molecular dynamics simulation. If C is smaller than 0.35, the method can overestimate its value if the number of configurations is not increased, especially for larger groups.
J Comput Chem 20: 1571–1576, 1999

Keywords: biopolymers; canonical correlation analysis; concerted motions; molecular dynamics simulation; statistical accuracy

Correspondence to: D. Genest; e-mail: genest@cnrs-orleans.fr

CCC 0192-8651 / 99 / 141571-06

Introduction

The flexibility of biopolymers is essential for their activity, and overall, concerted motions of groups of atoms play certainly a major role. In the last few years the canonical correlation analysis of data[1,2] applied to Molecular Dynamics simulation (MD) trajectories was used to quantify the level of correlation between the displacements of two groups of atoms in biopolymers such as oligonucleotides[3–6] or proteins.[7–9]. The canonical correlation analysis, first developed by Hotteling,[1] is a general method allowing to determine how two sets of random variates with centered and normalized Gaussian distribution are related. It has been used to describe the molecular internal dynamics of biomolecules as motions of rigid domains. Roughly, it consists on considering the time series obtained by MD simulation of two sets of linearly independent variables associated to two groups of atoms. The variables can be either the normalized fluctuations of atomic coordinates or of rigid body coordinates, which are sampled by the N values stored during the simulation. A single canonical correlation coefficient M (a scalar) is computed (using the home-made program TECOR), which gives an estimation of the global correlation between the two sets of variates considered as two objects. M reflects only a pure correlation, and is not biased by the direction of the atomic displacements as in other methods.[10,11] However, the question of the accuracy on the determination of this estimation was raised[4,5] but not yet solved. It was reported that the quality of the estimation depends on N, but also on the actual level of correlation and certainly on the number of variates in the two sets. Hotteling[1] has demonstrated that the standard error tends to an asymptotic behaviour in $N^{-1/2}$ when N is sufficiently large, but pointed out that there was no means of determining how large N must be. The present work addresses the influence of these different parameters on the accuracy of the computed correlation coefficient. A numerical study is presented in which samples of two sets of random variates representing coordinate fluctuations are first generated with a known global canonical correlation coefficient C between both groups. Then, the two sets of variates are analyzed with TECOR, which estimates this coefficient. The comparison between the actual value C and the estimated value M is presented for different values of C, of the sample size N and of the numbers of variates in the two sets.

Of course, the results of this study are not only valid for MD simulation problems, but for any purpose for which a level of correlation between two statistical objects defined by several variates is needed.

Methods

CANONICAL CORRELATION ANALYSIS OF DATA

Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ be two sets of n and m random variates, respectively, with Gaussian distributions. Let us consider $m \geq n$ for clarity. The variates are sampled by N values and are assumed to be normalized and centered. Each variate is, therefore, a N-dimensional vector $\mathbf{x}_i$ or $\mathbf{y}_j$, and X and Y define n- and m-dimensional subspaces of $\mathbf{R}^N$, respectively, provided the vectors in each subspace are linearly independent, which is the only condition for the canonical correlation analysis to be applied. The canonical correlation analysis consists on finding how these two subspaces are equivalent. Hotteling[1] has demonstrated that it is always possible to find orthogonal basis sets of unit vectors $\mathbf{a}_i$ $(i = 1, \ldots, n)$ and $\mathbf{b}_j$ $(j = 1, \ldots, m)$ in X and Y, respectively, such that:

$$\mathbf{a}_i \cdot \mathbf{b}_j = \delta_{ij} C_i \quad \text{for } j \leq n \qquad (-1 \leq C_i \leq +1)$$

$$\mathbf{a}_i \cdot \mathbf{b}_j = 0 \qquad \text{for } j > n$$

The initial vectors $\mathbf{x}_i$ and $\mathbf{y}_j$ are, therefore, linear combinations of the $\mathbf{a}_i$s and $\mathbf{b}_j$s, respectively. The $\mathbf{a}_i$s and $\mathbf{b}_j$s are called the canonical variates, and the C_is $(i = 1, \ldots, n)$ are their corresponding canonical correlation coefficients. An indicator of the correlation between both subspaces must be an invariant under internal linear transformation in X and in Y. Hotteling has shown that the only invariants of the system under such transformations are the C_is and functions of these quantities.[1] He proposed to use $Q = \prod_{i=1,n} C_i$ has an indicator of the correlation between X and Y. This indicator has the disadvantage that if only one C_i is zero, Q vanishes whatever the other C_is are. Later on, other indicators have been proposed such as the quadratic average of the C_is[12] defined as $C = \sqrt{(1/n)\sum_{i=1}^{n} C_i^2}$. If $C = 1$, all the C_is are equal to 1,

and the two subspaces are identical, meaning a full correlation between X and Y. On the contrary, if $C = 0$, all the C_is are zero, and both subspaces are orthogonal so that X and Y are fully uncorrelated. In the present work, as in previous ones,[3-9] C is used as an indicator of the correlation between X and Y, and is called global canonical correlation coefficient. Let $[X]$ (size N^*n) and $[Y]$ size (N^*m) be rectangular matrices, the ij element of which are the ith component of $\mathbf{x}_j$ or $\mathbf{y}_j$, respectively. Let $R_{11} = {}^T[X][X]$ (size n^*n) be the square symmetrical correlation matrix associated to the variates of X and $R_{22} = {}^T[Y][Y]$ (size m^*m) the equivalent one for the variates of Y. Let $R_{12} = {}^T[X][Y]$ be the rectangular correlation matrix mixing the $\mathbf{x}_i$s and the $\mathbf{y}_j$s (size n^*m) and R_{21} its transpose (size m^*n). It can be shown that the eigenvalues of the resulting n^*n matrix $R = R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$ are the C_i^2s. Thus, C can be easily calculated from the trace of R, without explicitly calculating the canonical vectors or the different eigenvalues of R.

ESTIMATION *M* OF THE GLOBAL CANONICAL CORRELATION COEFFICIENT

For application to MD trajectories data, one points out that the coordinates are generally not centered and not normalized. This has to be achieved prior to analyzing correlated motions.

The whole process used in TECOR to estimate C is schematized on Figure 1. The $x_i(k)$s and $y_j(k)$s are the values of the coordinates related to both groups of atoms, respectively, at configuration k ($k = 1, \ldots, N$). Each coordinate is averaged over the configurations, and its normalized deviation is calculated. These deviations are subsequently arranged to constitute two N rows matrices $[X]$ and $[Y]$ with n and m columns, respectively, rows corresponding to configuration and columns to variates. An estimation of the four correlation matrices R_{11}, R_{22}, R_{12}, and R_{21} is obtained from $[X]$,$[Y]$ and their transposes. After inversion of the estimated R_{11} and R_{22} matrices, the product matrix R mentioned above is calculated, and the estimated value M of the global canonical correlation coefficient C is obtained.

Different sources of errors can be mentioned. First, the numerical inversion of R_{11} and R_{22} could affect the accuracy. Second, the accuracy on the correlation matrices coefficients depends on the quality of the sampling. This is mainly this last point that is addressed in this study.

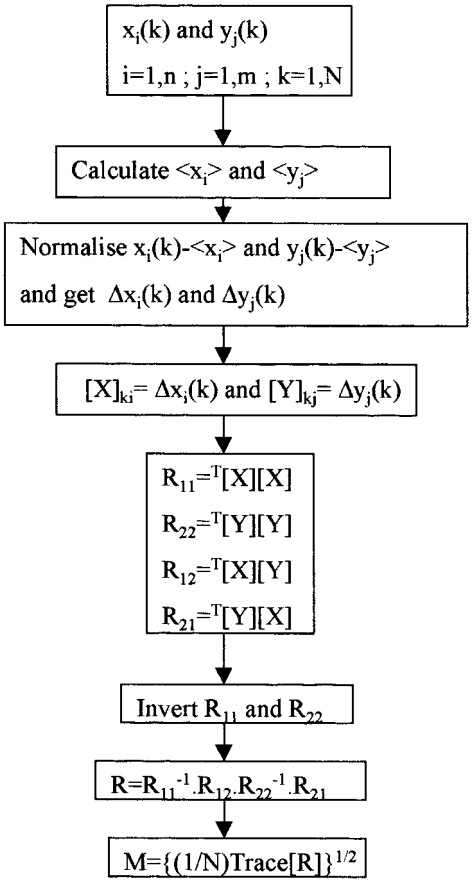

FIGURE 1. Overview of the method for calculating the global canonical correlation coefficient M between two sets of variables x_i and y_j sampled by N values. The averages are over the N values and $\Delta x_i(k)$ and $\Delta y_j(k)$ are normalized deviations. The exponent T in front of a matrix correspond to its transpose. The other symbols are described in the text.

GENERATION OF THE SAMPLES

For two groups of n and m variates, respectively, $(n + m)N$ random numbers are randomly picked up in a Gaussian distribution with 0 mean value and unit standard deviation, N being the size of the samples. Then they are split into $(n + m)$ series of N random numbers, the n first series corresponding to a first group, the other m series to a second group. Let $\{a_i\}$ $(i = 1, \ldots, n)$ the variates of the first group and $\{u_i\}$ $(i = 1, \ldots, m)$ those of the second group. At this point one has:

$$\langle a_i a_j \rangle = \delta_{ij}$$
$$\langle a_i u_j \rangle = 0 \qquad \forall\ i, j$$
$$\langle u_i u_j \rangle = \delta_{ij}$$
$$\langle a_i \rangle = \langle u_i \rangle = 0$$

m new series are subsequently generated according to:

$$b_i(k) = C_i a_i(k) + \left\{(1 - C_i^2)\right\}^{1/2} u_i(k) \quad \text{for } i \leq n$$
$$b_i(k) = u_i(k) \quad \text{for } n < i \leq m$$

where $a_i(k)$, $b_i(k)$, and $u_i(k)$ are the kth elements of the series a_i, b_i, and u_i, respectively, and the C_is are *a priori* chosen constants belonging to $[-1, +1]$. In these condition the a_i and b_i series verify:

$$\langle a_i b_j \rangle = \delta_{ij} C_i \quad \text{for } i, j \leq n$$
$$\langle a_i b_j \rangle = 0 \quad \text{for } j > n$$
$$\langle b_i b_j \rangle = \delta_{ij}$$
$$\langle b_i \rangle = 0$$

Thus, the a_is and b_is can be considered as samples of two groups of canonical variates as defined above, with canonical coefficients C_i. Let $\mathbf{a}_i$ and $\mathbf{b}_i$ be the corresponding canonical N-dimensional vectors. The vectors $\mathbf{x}_i$ and $\mathbf{y}_i$ to be analyzed are generally not canonical vectors and can be generated as :

$$\mathbf{x}_i = [A] \cdot \mathbf{a}_i$$
$$\mathbf{y}_i = [B] \cdot \mathbf{b}_i$$

where $[A]$ and $[B]$ are two any square matrices of size n^*n and m^*m, respectively, the elements of which are randomly picked up in a uniform distribution $[-\alpha, +\beta]$. In the present study, $\alpha = \beta = 5$. The two groups of vectors $\mathbf{x}_i$ and $\mathbf{y}_i$ can then be viewed as two groups of nonnormalized coordinate deviations. Their actual global canonical correlation coefficient C is known from the C_is, and its estimation M can be obtained from TECOR.

It is obvious that M depends on N, decaying from $M = 1$ for $N = 1$ to $M = C$ for $N \to \infty$. The rate at which it decreases must depend on n, m, and possibly on C. Because N is necessarily finite, M must also depends on the sample. To overcome this last dependence, all the results presented in this study correspond to an average over 10 samples for each set of the parameters N, m, and C. A larger averaging does not lead to a better accuracy.

In the present work, values of n and m were 3, 10, 20, 40, and 60, corresponding approximately to groups of 1, 3, 7, 13, and 20 atoms, respectively. C was taken between 0.01 and 0.850, either by setting all the C_is to C or by combining different values of C_is. The size N of the samples was varied from 500 to 5500 by steps of 500, covering the number of configurations usually stored during a MD simulation in the field of Biophysics.

Results and Discussion

Preliminary analysis showed that the estimated value M of the canonical correlation coefficient is not affected by using single or double precision for matrix inversion by the Gauss–Jordan elimination method.[13] Thus, numerical matrix inversion is not a cause of inaccuracy. It has also been verified that it does not depend neither on the $[A]$ and $[B]$ matrices used for deriving the $\mathbf{x}_i$s and $\mathbf{y}_i$s from the $\mathbf{a}_i$s and $\mathbf{b}_i$s.

Concerning the number of variates in the groups, the unexpected finding that M depends only on m (number of variates of the larger group) and not on n (number of variates of the smaller group) is obtained. This is illustrated in Table I for $m = 40$ and two values of n (3 and 40) for different N and different C.

Figures 2 and 3 show M as a function of N and C, respectively, for different m. It can be observed that M is practically equal to C for every m, when C is sufficiently high (≥ 0.35) and $N \geq 2000$, so that within these limits the estimation is correct. On the contrary, when C is below, the correlation is overestimated if N is too small, especially for large m. An extreme example for $N = 2000$ is given in the case $C = 0.01$ and $m = 60$, for which $M = 0.17$. Figure 2 shows that in some cases the correct value is not yet reached even for $N = 5500$.

These results allow a critical point of view on canonical time correlation functions calculated

TABLE I.
Estimation M of the Correlation Coefficient for $m = 40$ (m = Size of the Largest Group), and for Different Values of the Sampling Size N, and of the Actual Canonical Correlation Coefficient C.

	$C = 0.500$	$C = 0.250$	$C = 0.050$
$N = 500$	0.542 (0.020)	0.367 (0.016)	0.286 (0.025)
	0.540 (0.020)	0.367 (0.017)	0.290 (0.023)
$N = 1500$	0.515 (0.017)	0.292 (0.021)	0.170 (0.014)
	0.519 (0.017)	0.292 (0.020)	0.169 (0.015)
$N = 5000$	0.504 (0.010)	0.263 (0.017)	0.100 (0.011)
	0.508 (0.012)	0.266 (0.017)	0.099 (0.010)

M is the average over 10 different samples, and the maximum deviation is given in parenthesis. For each pair N, C two values are given corresponding to $n = 40$ (first line) and $n = 3$ (second line), n being the size of the smallest group.

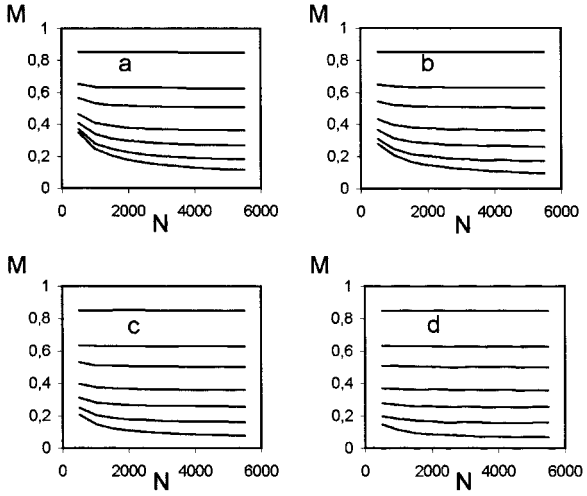

FIGURE 2. Estimation M of the global canonical correlation coefficient between two sets of m variables as a function of the sampling size N. (a) $m = 60$; (b) $m = 40$; (c) $m = 20$; (d) $m = 10$. On each plot are represented the curves obtained for different values of the actual correlation coefficient C (from top to bottom: 0.85, 0.63, 0.5, 0.35, 0.25, 0.15, 0.05).

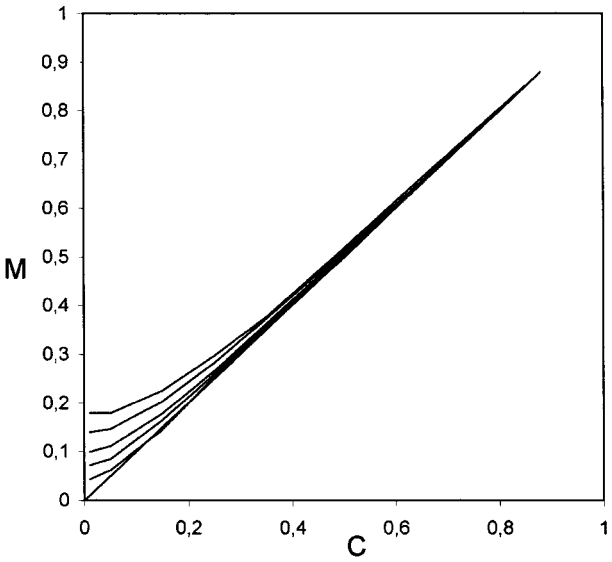

FIGURE 3. Estimation M of the global canonical correlation coefficient between two sets of m variables as function of the actual value C for a sampling size $N = 2000$. From top to bottom: $m = 60, 40, 20, 10, 3$, and the curve $M = C$.

from MD simulations.[5] Indeed, as time increases, the discrete points of such curves are usually computed with less and less configurations on one hand, and the correlation decreases on the other hand. These two facts add up to diminish the accuracy of the calculated correlation functions at a long period of time. As a consequence, for large groups of variates, only the first part of the curves can be properly analyzed, while the last part exhibits certainly too slow a decay.

The results presented above correspond to averages over 10 samples as, explained in the Method section, and M obtained from a given set of N, m, and C values is, thus, expected to be independent on the sampling. In the real case of performing MD simulations on large systems such as biopolymers, generally a single sample is available, and M obtained with TECOR can depend on the sampling, leading to an additional cause of inaccuracy for the estimation of C. The amplitude of the corresponding error can be estimated from our analysis, by measuring the dispersion of M over the 10 individual samples. It turns out that this dispersion is rather small, always smaller than 0.01, whatever the values of m, N, or C (Table I), which proves that TECOR converges rather well. However, one must have in mind that a bad sampling (i.e., a wrong simulation) can, of course, severely alter the statistics.

All the present analysis is valid as long as the variates verify Gaussian distributions such as, for example, coordinates fluctuating around a mean position. This is often the case for variables describing the equilibrium internal dynamics of well-structured biopolymers. However, when transition between different states are observed, the quantification of correlation becomes difficult because the number of these transitions is generally too small during the time course of the MD simulation to give good statistics. In particular, the data points of a time series obtained from MD simulation are generally time correlated, but this does not entail the validity of our approach as long as Gaussian distributions are observed, which is the case if the time course of the simulation is much larger than correlation times.

Conclusion

The present study is devoted to the level of confidence one can expect when the global correlation between two groups of variates is to be determined. Special attention is made to the case of analyzing correlated internal motions of groups of atoms inside a biopolymer, starting from MD simulation trajectories. In this case, the sample size is commonly in the range 1000–5000, and we have

limited our study to groups no larger than about 20 atoms. Within these conditions it is found that for sufficiently large correlation coefficients ($C \geq 0.35$) a correct value is always obtained from TECOR, provided 2000 configurations at least are stored during the simulation. If one is interested on the correlation between the displacements of two particular atoms or between the overall translations or overall rotations of two groups of atoms (two groups of three variables), a sampling size of 1000 configuration is sufficient, even for a small correlation. For larger groups of variates with a weak correlation ($C \leq 0.35$), a correct value can be reached only if the number of available configuration is sufficiently large; otherwise, C is overestimated. It, thus, appears that in practical cases the level of confidence on the estimated value M may be roughly classified into at least three categories (1) for $M \geq 0.35$, M is close to the actual value, (2) for $M \leq 0.20$, the groups are certainly almost uncorrelated, and M could even overestimate the correct value, and (3) for intermediate values ($0.20 \leq M \leq 0.35$), one has to be aware that large groups could seem to be moderately correlated, while in reality, they could be practically not. This study also shows that a sufficiently large number of configurations has to be stored during an MD simulation when correlated motions have to be analyzed.

Acknowledgments

I am grateful to A. Boyer for his technical assistance in programming the generation of the samples.

References

1. Hotteling, H. Biometrika 1936, 28, 21.
2. Girshick, M. A. Ann Mathemat Stat 1939, 10, 203.
3. Briki, F.; Genest, D. Biophys Chem 1994, 52, 35.
4. Briki, F.; Genest, D. J Biomol. Struct Dynam 1995, 12, 1063.
5. Genest, D. Biopolymers 1996, 38, 389.
6. Genest, D. Eur Biophys J 1998, 27, 283.
7. Hery, S.; Genest, D.; Smith, J. C. Phys B 1997, 234–236, 175.
8. Hery, S.; Genest, D.; Smith, J. C. J Chem Inf Comput Sci 1997, 37, 1011.
9. Hery, S.; Genest, D.; Smith, J. C. J Mol Biol 1998, 279, 303.
10. Ichiye, T.; Karplus, M. Proteins Struct Funct Genet 1991, 11, 205.
11. Hünenberger, P.; Mark, A.; van Gunsteren, W. J Mol Biol 1995, 252, 492.
12. Pontier, J.; Dufour, A. B.; Normand, M. Le Modèle Euclidien en Analyse de Données; Université de Bruxelles: Bruxelles, 1990.
13. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. Numerical Recipes: The Art of Scientific Computing; Cambridge University Press: Cambridge, 1986.